# Vector Probability Diffusion

### Alvaro Pardo[*]    Guillermo Sapiro[†]

October 8, 1999

### Abstract

A method for isotropic and anisotropic diffusion of vector probabilities in general, and posterior probabilities in particular, is introduced. The technique is based on diffusing via coupled partial differential equations restricted to the semi-hyperplane corresponding to probability functions. Both the partial differential equations and their corresponding numerical implementation guarantee that the vector remains a probability vector, having all its components positive and adding to one. Applying the method to posterior probabilities in classification problems, spatial and contextual coherence is introduced before the MAP decision, thereby improving the classification results.

## 1   Introduction

The basic motivation of this work is to introduce contextual information into image segmentation tasks by adding spatial coherence to the posterior probabilities corresponding to the classes present in the scene. In [9] the authors proposed to introduce this spatial coherence via anisotropic diffusion of the posterior probabilities. In this method, posterior probabilities for each class are computed, they are then independently anisotropically smoothed [5], and only then the MAP decision rule

is applied. Connections between this method and other popular techniques like Markov Random Fields and Continuous Relaxation Labeling were shown as well. Although this technique was already found to be useful for problems like MRI and SAR segmentation, two main difficulties are encountered in this work. First, each posterior probability (one per class) is independently diffused, thereby ignoring the intrinsic correlation between them. Second, due to the independent processing, the posterior probabilities are not guaranteed to add to one even after a very short diffusion time. To overcome this, the authors normalize the posterior probabilities after each discrete iteration, normalization that has a non-trivial effect on the diffusion process itself.

In this work we show how to isotropically and anisotropically diffuse a vector of probabilities while preserving the diffusion on the probabilities semi-hyperplane. In other words, the diffusion process guarantees that the vector components are positive and add to one. Moreover, the numerical implementation of this system of coupled PDE's also preserves these properties, thereby removing the necessity to project back into the semi-hyperplane. This approach then overcomes both difficulties mentioned above and can be directly incorporated in the segmentation technique, replacing the component-by-component diffusion. Although the work is here discussed in the framework of posterior diffusion for image classification, it is clear that the technique can be applied to the diffusion of other probabilities in other applications as well.

The basic idea of our approach is then to keep the diffusion process in the semi-hyperplane corresponding to the vector of probabilities. This can be seen as a particular case of diffusion on general manifolds, topic that has recently received a considerable amount of attention from the image processing and computer vision communities [1, 4, 7, 8]. That is, instead of having the data defined on $\mathbb{R}^n$, it is defined on a general manifold $N$ (e.g., an hypersphere for directional data [1, 4, 8] or a semi-hyperplane for probabilities). In addition, as we will see below, the results here introduced can be obtained investigating the basic invariants of diffusion processes.

## 2    Problem Formulation and Basic Equations

Assume that a vector of a-posteriori probabilities $p$, mapping the image domain $\Omega$ in $\mathbb{R}^2$ to the manifold $\mathcal{P} = \{p \in \mathbb{R}^m : \|p\|_1 = 1, p_i \geq 0\}$, is given. Each component $p_i$ of $p$ equals the posterior probability of a class $c_i \in \mathcal{C} = \{c_i : i = 1, ..., m\}$. These posterior probabilities can be obtained for example via Bayes rule.

If we view the vectors $p$ as a vector field, one possible way to add spatial coherence into the classification process is to diffuse the distance between points in $\mathcal{P}$, propagating the information in the probability space, before a pixelwise MAP decision is made. Inspired by the harmonic maps theory [8], the "distance" between two differential adjacent points in $\mathcal{P}$ depends on $\|\nabla p\| := \sqrt{\sum_{i=1}^{m} \|\nabla p_i\|^2}$. This is the *gradient of the probability vector*. Giving a function $\rho : \mathbb{R} \to \mathbb{R}$ (we will later discuss different selections of $\rho$), we proceed to solve the following minimization process:

$$\min_{p \in \mathcal{P}} J_\rho, \quad J_\rho := \int_\Omega \rho(\|\nabla p\|) d\Omega. \tag{1}$$

Note that the minimization is restricted to the semi-hyperplane $\mathcal{P}$.

From this, the system of coupled diffusion equations is obtained via the gradient descent flow corresponding to this energy.

**Property 1** *The gradient descent of $J_\rho$ restricted to $\mathcal{P}$ is given by*

$$\frac{\partial p}{\partial t} = \nabla . \left( \frac{\rho'(\|\nabla p\|)}{\|\nabla p\|} \nabla p \right) .^1 \tag{2}$$

*Proof:* This is easily obtained computing the Euler-Lagrange and projecting it into $\mathcal{P}$. □

Note again that the minimization is performed in the probability space $\mathcal{P}$ (a semi-hyperplane), and the system of equations (2) guarantees that $p(t) \in \mathcal{P}$ for all $t$. We have therefore obtained a system of coupled PDE's that preserves the unit $L_1$ norm which is characteristic of probability vectors (the components $p_i(t)$ are positive as well).

It can be shown that $\sum_{i=1}^{m} \|\nabla p_i\|^2 = 2\|\nabla p_1\|^2 - 2\sum_{i=2}^{m} \sum_{j=i+1}^{m} \nabla p_i . \nabla p_j$. Therefore, when the number of classes is $m = 2$, $\|\nabla p\|^2 = 2\|\nabla p_i\|^2$, $i = 1, 2$, and the method is equivalent to separately applying the diffusion to each posterior. For $m \geq 3$, the second term is like the correlation between different components of $p$. This term is not present if we diffuse each posterior probability on its own. This coupling is important to improve the classification results.

If we select $\rho$ as the $L_2$ norm, $\rho(x) = x^2$, then (2) becomes the well known linear heat equation $\frac{\partial p}{\partial t} = \nabla^2 p$, which isotropically diffuses $p$. It is interesting to note that the heat equation that was previously used to denoise signals as well to generate the so called scale spaces, preserves the diffusion in the probability semi-hyperplane $\mathcal{P}$. This is expected, since it is well known that this equation holds the maximum principle and preserves linear combinations.

---

[1] This equations is an abbreviated notation for a set of PDE's of the form $\frac{\partial p_i}{\partial t} = \nabla . \left( \frac{\rho'(\|\nabla p\|)}{\|\nabla p\|} \nabla p_i \right)$, $i = 1, ..., m$.

The isotropic flow has no coupling between the probability components, and does not respect boundaries. Therefore, as classically done for scalar diffusion [6], a more robust norm is selected. For example, we can select the $L_1$ norm, $\rho(x) = |x|$, obtaining

$$\frac{\partial p}{\partial t} = \nabla . \left( \frac{\nabla p}{\|\nabla p\|} \right), \tag{3}$$

clearly anisotropic, with a conduction coefficient controlled by $\|\nabla p\|$. Once again, from the basic invariants of this flow, the preservation of the vector on the $\mathcal{P}$ space was expected.

## 2.1   Numerical Implementation

The vector probability diffusion equations are to be implemented on a square lattice. The vector field $p$ is then $p_{j,k}$, where $j, k$ is the position in the lattice. (For simplicity, we do not write the subscript $i$ indicating the probability components.) First we rewrite the equation (2) in a simpler way, $\frac{\partial p}{\partial t} = \nabla . (g \nabla p)$, $g = \rho'(\|\nabla p\|)/\|\nabla p\|$, and then apply a standard numerical scheme (assuming $\Delta x = \Delta y = 1$):

$$\begin{aligned}
\frac{p_{j,k}^{t+\Delta t} - p_{j,k}^t}{\Delta t} = g_{j+\frac{1}{2},k}(p_{j+1,k}^t - p_{j,k}^t) &\quad - \quad g_{j-\frac{1}{2},k}(p_{j,k}^t - p_{j-1,k}^t) + \\
g_{j,k+\frac{1}{2}}(p_{j,k+1}^t - p_{j,k}^t) &\quad - \quad g_{j,k-\frac{1}{2}}(p_{j,k}^t - p_{j,k-1}^t).
\end{aligned} \tag{4}$$

The condition on $\Delta t$ for stability is easily computed to be

$$\Delta t \leq \frac{1}{4 \max_{j,k}\{g_{j+\frac{1}{2},k}, g_{j-\frac{1}{2},k}, g_{j,k+\frac{1}{2}}, g_{j,k-\frac{1}{2}}\}}. \tag{5}$$

To allow the existence of discontinuities in the solution we use the approximation of the gradient developed in [3], $\|\nabla p\|_{i,j} = \sqrt{\|D_{x_i}^+ p_{j,k}^t\|^2 + \|m(D_{x_j}^+ p_{j,k}^t, D_{x_j}^- p_{j,k}^t p_{j,k}^t)\|^2}$, where $D_{x_i}^+$ and $D_{x_i}^-$ are the forward and backward differences and $m(x,y)$ is the minmod operator applied to each component of $p$, $m(x,y) = \mathrm{minmod}(x,y) = \min(|x|,|y|)(\mathrm{sign}(x) + \mathrm{sign}(y))/2$.

**Property 2** *If $\Delta t$ fulfills the stability condition (5) then $p_{j,k}^{t+\Delta t} \geq 0$   and   $\|p_{j,k}^{t+\Delta t}\|_1 = 1$, so the evolution given by (4) lives always in the manifold $\mathcal{P}$. Moreover, if $p_m^t := \min\{p_{j+1,k}^t, p_{j-1,k}^t, p_{j,k+1}^t, p_{j,k-1}^t\}$ and $p_M^t := \max\{p_{j+1,k}^t, p_{j-1,k}^t, p_{j,k+1}^t, p_{j,k-1}^t\}$, the solution satisfies a maximum (minimum) principle: $p_m^t \leq p_{j,k}^{t+\Delta t} \leq p_M^t$.*

4

We then conclude that also the discrete equation (4) lives in the manifold $\mathcal{P}$, and there is no need for a projection back into the semi-hyperplane, in contrast with the approach by Teo *et al.*, as discussed in the introduction.[2]

# 3  Examples

We now present examples of the vector probability diffusion approach here presented applied to image segmentation, following the technique introduced in [9] . First, posterior probabilities are computed for each class using Bayes rule. These are diffused with the anisotropic vector probability diffusion flow (instead of the scalar one used in [9]), and then the MAP decision rule is applied. We will compare our results with those obtained by Teo *et al.* in order to show the improvements obtained with the vectorial approach.

The first example is based on synthetic data. The original image, containing four classes, is shown on the top-left of Figure 1. We then add Gaussian noise, and the objective is to segment the image back into four classes. The figure then shows the classification results without posterior diffusion, followed by the classification results corresponding to two time steps for the vector posterior diffusion approach and two time steps for the scalar approach. Using the correlation between the posterior probabilities, as done by the vector approach here introduced and not by the scalar one in [9], is of particular importance when one of the classes has less weight than the others. As shown by this example, this class (white dots) will be mostly missed by the scalar approach. Comparing the classification errors for class 4 (white dots) in Table 1, we see that the lowest classification error is with vector probability diffusion, 15.82% against 18.46% for the scalar approach. Furthermore, the classification error is more stable with the vectorial approach, and for the same average error (approx. 10%), the lowest classification error for class 4 is obtained with our method, 26.95% against 61.04% (see Table 1).

Figure 2 shows the results of our approach applied to the segmentation of video SAR data.

---

[2]To complete the implementation details, we need to address possible problems when $\|\nabla p\|$ vanishes or becomes very small. As it is standard in the literature, we define $\|\nabla p\|_\beta = \sqrt{\beta^2 + \|\nabla p\|^2}$, and use this instead of the traditional gradient. (The stability condition (5) becomes $\Delta t \leq \beta/4$.) To select the value of $\beta$ we propose to look at $\beta$ as a lower diffusion scale, since in the discrete case probability differences lower than $\beta$ will be diffused with a conduction coefficient approximately inverse to $\beta$. We set the value of the lower scale $\beta$ in the range $[0.001, 0.01]$.
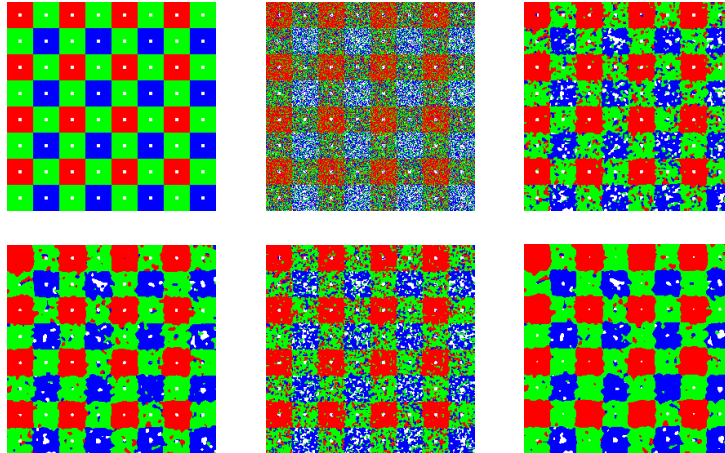
Figure 1: Left to right and top to down: Original image, classification without diffusion, results of vector probability diffusion for 18 and 25 iterations respectively, and results of the scalar approach for 10 and 15 iterations.

Three classes are considered, {shadow, object, background}, each one modeled as a Gaussian distributions. [3] The first image in the sequence is segmented by hand to obtain an estimation of the parameters of the distributions. For the rest of the images we perform the following process: (a) The diffused posteriors of the previous image are used as priors and after 30 iterations the new means and variances are estimated; (b) To obtain a better segmentation which does not depend on the previous priors, a new diffusion is applied using uniform priors. In this way the small details are captured; (c) The steps (a) and (b) are iterated and in the last iteration the idea of (a) is used.

As we can see in the Figure 2 the results are very good, and the small details are captured. These results outperform those reported in [2].

## 4    Conclusions

We have presented a method for the diffusion of probability vectors. Both the corresponding diffusion equations and their numerical implementation guarantee that the diffused quantities remain probabilities, that is, are positive and add to one. This method can be used to incorporate spatial coherence into MAP-based segmentation techniques. We have shown that the approach here

---

[3]If we learn the distribution using a kernel method we find that two of the classes are nearly Gaussian and the last one could be approximated with a Gaussian.

| Iter. | Class 1 | Class 2 | Class 3 | Class 4 | Average |
|---|---|---|---|---|---|
| 0 | 53.77 | 26.67 | 53.68 | 27.34 | 46.65 |
| 10 | 47.63 | 17.47 | 47.40 | 22.46 | 39.70 |
| 15 | 33.26 | 4.76 | 33.30 | 17.77 | 26.02 |
| 18 | 24.01 | 2.33 | 24.83 | 15.82 | 18.95 |
| 20 | 19.75 | 1.50 | 20.88 | 17.77 | 15.78 |
| 25 | 12.16 | 0.81 | 14.39 | 26.95 | 10.70 |
| 5 | 53.70 | 26.51 | 53.61 | 27.34 | 46.55 |
| 7 | 50.02 | 18.74 | 49.85 | 24.90 | 41.85 |
| 10 | 35.06 | 6.24 | 35.03 | 18.46 | 27.69 |
| 15 | 17.48 | 1.51 | 18.99 | 24.32 | 14.40 |
| 20 | 18.11 | 0.76 | 13.14 | 61.04 | 10.34 |

Table 1: Classification errors per class and average for the example in Figure 1. The first row corresponds to the classical MAP approach, where no diffusion has been applied to the posterior probabilities. The next four rows correspond to four different time steps of the vector probability diffusion approach here presented, followed by four time steps for the scalar approach of Teo *et al.*.

introduced outperforms similar techniques that independently diffuse each probability component, adding the constraints as an additional, not formally justified, projection step.

# References

[1] T. Chan and J. Shen, "Variational restoration of non-flat image features: Models and algorithms," *UCLA CAM-TR 99-20*, June 1999.

[2] S. Haker, G. Sapiro, and A. Tannenbaum, "Knowledge-based segmentation of SAR data with learned priors," *IEEE Trans. Image Processing*, to appear.

[3] S. Osher and L. Rudin, "Featured-oriented image enhancement using shock filters," *SIAM Journal of Numerical Analysis* **27:4**, pp. 919–940, August 1990.

[4] P. Perona, "Orientation diffusion," *IEEE Trans. Image Processing* **7**, pp. 457-467, 1998.

[5] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE-PAMI* **12**, pp. 629-639, 1990.
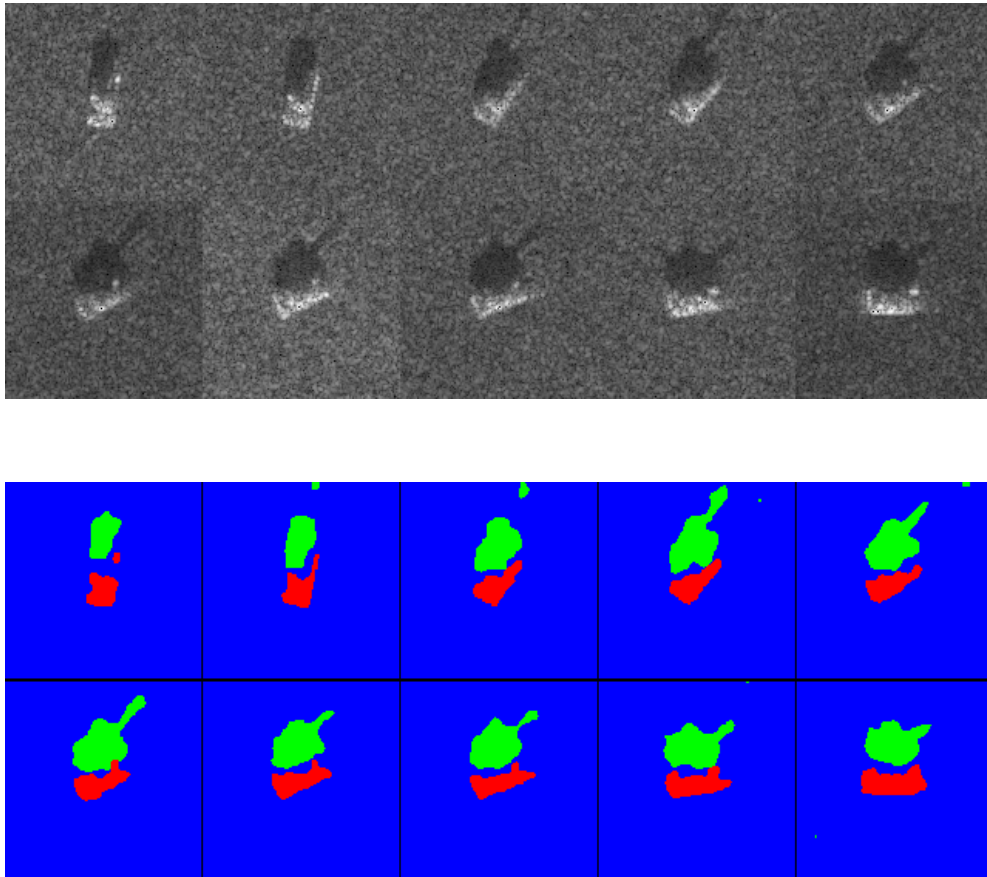
Figure 2: Original images and classification results with MAP vector probability diffusion.

[6] L. Rudin, S. Osher, and E. Fatemi, "Nonliner total variation based noise removal algorithms," *Physica D* **60**, pp. 259–268, 1992.

[7] N. Sochen, R. Kimmel, and R, Malladi, "A general framework for low level vision," *IEEE Trans. Image Processing* **7**, pp. 310-318, 1998.

[8] B. Tang, G. Sapiro, and V. Caselles, "Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case," *International Journal Computer Vision*, to appear.

[9] P. Teo, G. Sapiro, and B. Wandell, "Creating connected representations of cortical gray matter for functional MRI visualization," *IEEE Transactions on Medical Imaging* **16:6**, pp. 852–863, December 1997.